Accepted Manuscript

How B-DNA Dynamics Decipher Sequence-Selective Protein Recognition

Federica Battistini, Adam Hospital, Diana Buitrago, Diego Gallego, Pablo D. Dans, Josep Lluis Gelpí, Modesto Orozco

PII: S0022-2836(19)30451-6
DOI: https://doi.org/10.1016/j.jmb.2019.07.021
Reference: YJMBI 66219

To appear in: *Journal of Molecular Biology*

Received date: 13 May 2019
Revised date: 9 July 2019
Accepted date: 10 July 2019

Please cite this article as: F. Battistini, A. Hospital, D. Buitrago, et al., How B-DNA Dynamics Decipher Sequence-Selective Protein Recognition, Journal of Molecular Biology, https://doi.org/10.1016/j.jmb.2019.07.021

How B-DNA dynamics decipher sequence-selective protein recognition

Federica Battistini[1], Adam Hospital[1], Diana Buitrago[1],

Diego Gallego[1], Pablo D. Dans[1], Josep Lluis Gelpí [2] and Modesto Orozco[1,2,*]

[1] Institute for Research in Biomedicine (IRB Barcelona). The Barcelona Institute of Science and Technology, Barcelona, Spain.

[2] Department of Biochemistry and Biomedicine, University of Barcelona, Barcelona, Spain.

* To whom correspondence should be addressed: modesto.orozco@irbbarcelona.org

Abstract

The rules governing sequence-specific DNA-protein recognition are under a long-standing debate regarding the prevalence of *base* versus *shape* readout mechanisms to explain sequence specificity and of the *conformational selection* versus *induced fit* binding paradigms to explain binding-related conformational changes in DNA. Using a combination of atomistic simulations on a subset of representative sequences and mesoscopic simulations at the protein-DNA interactome level, we demonstrate the prevalence of the *shape readout* model in determining sequence-specificity and of the

conformational selection paradigm in defining the general mechanism for binding-related conformational changes in DNA. Our results suggest that the DNA uses a double mechanism to adapt its structure to the protein: it moves along the easiest deformation modes to approach the bioactive conformation, while final adjustments require localised rearrangements at the base pair step and backbone level. Our study highlights the large impact of B-DNA dynamics in modulating DNA-protein binding.

Keywords

DNA-protein recognition, Molecular dynamics, PDB data mining, Structural analysis, Principal component analysis.

Introduction

DNA-protein recognition, an essential step in gene regulation, depends on both the accessibility of the DNA and its intrinsic affinity for the protein. Accessibility is related to the chromatin fold and to the presence of competing proteins, while affinity is determined by the formation of protein-DNA contacts and by the cost of deforming the DNA duplex from the naked to the bound bioactive conformation. Two extreme situations can be envisioned in DNA-protein binding: one where the complex formation follows a base readout mechanism in which specific DNA-protein contacts determine sequence specificity, and another one where the binding follows an shape readout model, *i.e.* DNA deformability properties explain sequence-specific binding [1]. Shape (indirect) recognition describes protein-DNA recognition mechanisms that depend on the ability of a DNA sequence to adopt a conformation that facilitates its binding to the protein or that intrinsically has the matching conformation for the protein binding.

Protein-DNA shape recognition involves the formation of specific binding sites for positively charged amino-acids, ARG/LYS, indirect contacts with phosphates some direct hydrogen bonds established with DNA bases and interactions mediated through water molecules, depending on the different solvation, above all upon binding the release of water molecules from the protein-DNA interface providing a favorable entropic contribution and it is important for selectivity [2–4] Very often, protein binding leads to a conformational change in DNA, and again two different models can be proposed to explain the connection between structural flexibility and binding: conformational selection and induced fit. The recognition modes contribute to the overall protein-ligand binding mechanism that couples conformational selection and conformational changes, that depend on the ligand, in this case DNA, and protein properties and on multiple conditions, including the interactions between the biomolecules, their concentrations [5] and the rate of the conformational transition [6]. Within the conformational selection paradigm, the deformation energy required to move the DNA from naked to bioactive conformation is small, typically within the DNA thermal fluctuation and it is then sampled spontaneously in the "unbound" state. On the contrary, according to the induced fit model the DNA deformation energy required for binding is large, hampering the spontaneous population of the bound state from the naked B-DNA dynamics.

In the last decades, many experimental and computational studies analysed the specificity of the protein-DNA binding to better understand their recognition [7–15]. Thus, databases have been built and software has been developed to study the interactions, affinity and selectivity in protein-DNA binding [16]. Databases store, for example, preferred DNA binding sites for a large number of proteins as determined by

SELEX-seq/HT-SELEX, microarrays, chromatin immunoprecipitation and others [8–15,17,18]. Such sequence-based information is combined with the structural analysis of known protein-DNA complexes to derive interaction rules which are implemented in a variety of statistical methods (11–16). Alternatively, *ab initio* approaches to the study of protein-DNA interactions are based on the use of energy-based *in silico* methods, which use protein-DNA direct interaction terms [22], and deformation energies derived from DNA properties [23] to recognize binding sites through structural signals.

Despite the variety of experimental and computational studies on DNA-protein binding, the relative importance of base versus shape readout is unclear, and no consensus exists on the prevalence of induced fit or conformational selection paradigms [1]. Certainly, part of the problem is due to the discrepancy existing in the experimental information available, as data obtained from HT-SELEX [18,24], footprinting [25], protein binding microarrays (PBM) [8,26] or ChIP-chip/ChIP-seq experiments [12,27,28] depend on the experimental technique and conditions making statistical methods noisy and often over-trained to reproduce just one type of data. For this reason experimentally-trained statistical methods should be complemented with approaches based on the calculation of interaction and deformation energies by means of physical models, which are not influenced by the noise of high-throughput experimental data [29–39].

In this article we present an *in silico* analysis of the role of DNA conformational flexibility in the formation of protein-DNA complexes. The systematic evaluation of the conformational changes of physiological DNA associated to protein binding was performed using molecular dynamics simulations, with the newly-refined parmbsc1

force field which allowed performing analyses of DNA structure and flexibility with accuracy similar to that of current experimental techniques [40–44], and mesoscopic simulations, focusing on DNA sequence preferences. Results presented here provide convincing evidence for the impact of the shape readout on the DNA-protein interactome, and for the prevalence of conformational selection mechanism in defining binding-related conformational change in DNA, at least in those cases where the protein does not have a clear disruptive effect on the DNA structure. Our results suggest that DNA adapts to the presence of the interacting protein following a dual mechanism: global movements are facilitated as coded in the essential dynamics of the duplex, while local rearrangements are related to displacements at the base pair step level and are coupled to complex backbone rearrangements. In this analysis we took into account that the torsion angles from experimental data (NMR and X-ray) are difficult to determined and are not completely captured and validated (for a discussion on experimental backbone torsion angles reliability see references [45,46]) . However results presented here show how sequence-dependent B-DNA dynamics are a key player in modulating DNA-protein recognition and that dynamics of isolated DNA in physiological conditions is important in determining DNA-protein interaction, independently of the specific Protein-DNA binding motif.

Results and Discussion

MD simulations of the 50 naked DNA sequences (see Methods, Fig. 1) provided stable trajectories without any remarkable distortion after 500 ns of simulation time. The origin of the starting structure (canonical B-form or bound state) is not relevant (see Supplementary Fig. S2) supporting the idea that simulations are sampling equilibrium conformations without memory of the initial structure. The conformational space

sampled in the trajectories agrees very well with the one expected for B-DNA duplexes [40], leading to a set of structures that lost memory of the initial experimental ones (X-ray or NMR) [40,41,47].

How is the intrinsic geometry of the DNA modified by protein binding? The comparison between the experimental DNA structure and the conformational space sampled by the naked DNA in the MD simulations using Hoteling's statistics (see methods) revealed that in general the bound DNA structure falls within naked DNA conformational space (red circle in Fig. 2 for rise and roll and Supplementary Fig. S3 for the remaining bp parameters). In the few cases, where DNA ensembles show local differences from the bound DNA structure, we checked meticulously for potential uncertainties in the reported experimental structure. We found three NMR structures where doubts may exist regarding certain structural details. For example, 1C7U shows highly unusual rise, slide and roll values (see Supplementary Fig. S4), in regions away from the protein, signalling potential artefacts in the refinement leading to abrupt and compensatory helical profiles [41]. 1ZGW shows an unusual rise profile at the duplex termini (d(A_3A_4) and d($A_{15}A_{16}$)); the large rise in the latter may be explained by the partial intercalation of Phe_{114}, but the large and unusual rise (around 5 Å) at the other base pair step is very difficult to explain as there are no interacting protein residues in the vicinity (see Supplementary Fig. S5). Finally, 2STW shows further unusual rise values (see Supplementary Fig. S6) at the central d(T_7T_8) step, that can be explained by the presence of the partial intercalation of Tyr_{85}, and at the termini (d(C_2G_3) and d($C_{15}G_{16}$)) where the high rise is suspicious as it is not justified by any protein-DNA interactions. Complexes 1T9I, 2KDZ, 2L1G, 1MOW, 1YTF, 2QHB, 1YFI, 1AZP, 1A0A, 1CDW, 3F27 and 3U2B show some unusual values in helical parameters of the DNA (Fig.

2), that could however be explained by direct contacts with the protein. For example, partial intercalation explains the large kink in the last 3 structures (1CDW, 3F27 and 3U2B), while strong salt-bridge contacts of DNA backbone with cationic residues of the protein could explain unusual roll and rise values in 1A0A (see Fig. 3 and Supplementary Figs. S7-9).

What is the backbone conformation required for protein binding? In general, backbone angles in DNA-protein complexes remained in the conformational space sampled by naked DNA simulations, like the base pair parameters (Fig. 2). However, when large alterations in helical coordinates are required, they are achieved by concerted changes in the backbone angles (α, β, χ, ε, γ, phase and ζ) [48,49]. Kinks, highly bent base pair steps, are linked to significant alterations in sugar puckering, which are rare in unperturbed DNA, and that can be coupled to other backbone changes. It seems that coordinated movements of α/γ and, to a lesser extent, ε/ζ are frequent in protein-distorted DNA and are related to the tendency of the phosphates to approach cationic residues in the interacting protein. Changes use to be localized at regions of direct contact with the protein. An example is given in Fig. 3 that shows a detailed analysis of PDB ID 1A0A, where distortions in both helical parameters and backbone angles are visible at regions interacting with protein. In particular, base pair steps with high roll (CC and CG) are correlated with unusual α/γ angle and are in contact with protein residues ARG, LYS and GLU (Fig. 3). We also detected correlations between distorted base pair step parameters and unusual backbone values for α/γ and phase (in kinked structures also χ and β) angles where the protein residues are in contact with the DNA in particular for the structures PDB ID 1CDW, 3U2B and 3F27 (Supplementary Figs. S7-S9). Overall, our analysis conclude that unbound DNA backbone is rather

flexible under physiological conditions, and there are few cases where unusual conformations of the backbone, not present in the naked ensemble, are required for adopting the bioactive conformation.

What is the energy cost of deforming helical coordinates for binding? The thermal energy fluctuation calculated for naked DNA along the MD, amounts to around 2.5 ± 0.1 kcal/mol·bp (see Methods) and, accordingly, when the energy cost of achieving the bound state is lower than this value, we can conclude that the bound state can be spontaneously sampled (being thermodynamically accessible at physiological conditions) by the naked DNA. Inside this energetic range the DNA-protein binding follows a behavior that falls into the conformational selection paradigm. In contrast, when distortion energy cost is larger than this value we can conclude that the DNA needs external effector to change structure and adopt the bioactive conformation, leading to the induced fit mechanism (Fig. 4a). We considered a margin of twice the free DNA fluctuation energy as a twilight zone (red area in Fig. 4a, between 2.5 and 5.0 kcal/mol·bp), where hypothetically both recognition modes, conformational selection and induced fit, coexist [50].

Mesoscopic calculations (Fig. 4a) indicate that for 33 of the 50 complexes considered, the energy cost for the DNA to adapt to the bioactive conformation is within the free DNA fluctuation energy range and fall inside the defined blue area (Fig. 4a). That is, in most cases, binding follows the requirements of the conformational selection mechanism. The induced fit mechanism explains binding in 12 of the 50 complexes (white area in Fig. 4a), while the remaining 5 cases can be labelled as in the twilight

zone (red area in Fig. 4a), where possibly both mechanisms contribute to the binding. Our results indicate that conformational selection seems to be at least twice more prevalent than induced fit in modulating DNA-protein binding in our set of representative DNA-protein complexes.

Are essential deformation modes coupled to protein-induced DNA deformation? Large protein-induced conformational transitions in DNA (initial RMSD between naked and bound structure ($RMSD_{in}$ > 5 Å, value to delineate the boundary between large and small DNA distortion, defined by the average plus one standard deviation of the dots in the blue area in Fig. 4a) are possible thanks to good alignment between the transition vector (from the naked to the bound structure) and the essential deformation (ED) modes of the naked DNA (see Methods). Such an alignment is not a necessary for small protein-induced transitions ($RMSD_{in}$ < 5 Å), where only local rearrangements are required. This is shown in the dependence of the $RMSD_{in}$ and the squared overlap calculated between ED modes and the transition vector, as well as in the correlation between $RMSD_{in}$ and the distance covered applying the essential deformation modes of the naked DNA (see Methods and Fig. 4b-c). Our results strongly suggest that DNA adapts to protein shape following a dual mechanism. On one hand, global deformations happen along preferred deformation modes and bring the naked DNA structure close to that of the bioactive (protein-bound) conformation (around 3-4 Å in RMSD), independently of the original $RMSD_{in}$ (Fig. 4d). On the other, small movements at the base pair step level are required for the fine grain adjustments to reach for the perfect complementarity between the protein and the DNA.

We found that in general, physiological B-DNA is flexible enough to easily sample its bioactive conformation without the presence of the protein, supporting the prevalence of the conformational selection over the induced fit mechanism, at least for protein-DNA complexes where the protein does not break the Watson-Crick base pairing. In those cases where reaching the bioactive conformation implies mild distortions, they typically involve local re-arrangements in the base-pair step geometry and small backbone changes. However, when the required distortion is large, DNA reaches the bioactive state by first moving along the low energy essential deformation modes, and finally by way of local rearrangements fine tuning DNA conformation sub-states.

What is the driving force for large protein-induced structural deformations? As discussed above, many of the complexes studied here require structural distortions in the DNA that are easy to achieve from the naked ensemble, in agreement with the conformational selection model. There are, however, a few complexes for which conformational changes come at a high deformation cost (see Fig. 4a) and we were intrigued on the driving force of these distortions. A detailed analysis of these cases show that the structural deformation induced by the protein results in changes in the electrostatic field of the DNA, which obey the need of DNA to accommodate to the protein interacting residues. Thus, upon binding, regions of the DNA facing apolar residues become less cation-philic, while negative potential is reinforced in those regions facing Arg/Lys-rich areas (see selected molecular interaction potential, MIP, maps in Fig. 5). Interestingly, in several cases the negative MIP regions detected by the probe and generated by the distortion of the DNA geometry coincide with sites occupied by positively charged protein residues, LYS/ARG, at the interaction interface. Changes in structure seem to create anchoring points for cationic residues in protein tails that

would otherwise be disordered. The analysis of the electrostatic surface of these 3 cases with different degrees of distortions, suggests a subtle protein-DNA structural interplay where the ordered part of the protein distorts the DNA towards the bioactive state, leading to changes in the DNA electrostatic potential, which in return generates additional anchoring points for the disordered protein tails.

What is the relative prevalence of conformational selection and induced fit binding modes? For each one of 174 complexes in the curated dataset representative of the entire DNA-protein interactome the mesoscopic deformation energy associated to binding was computed (see Methods and Supplementary Methods). Very interestingly, a vast majority of the cases follow a pure conformational selection binding mode (71%, blue area in Fig. 6), 18 % of the cases fall in the twilight zone and induced fit explains around 11% of the complexes (see Fig. 6), where extremely bent or even kinked DNA is obtained by direct protein-nucleotide contacts.

Considering that the structures selected are the 17% over the entire dataset of the no-redundant PDB protein-DNA structures, this 71% corresponds to a 24% in the entire repository.

Even potential bias derived from PDB composition cannot be ruled out, our results strongly support, for complexes where the B-DNA structures are not strongly altered by the protein (mismatch/broken/unpairing, Supplementary Fig. S1), the prevalence of the conformational selection model over the induced fit one, as anticipated by atomistic simulations on the 50 selected complexes. Interestingly these two group are also characterised by different binding specificity. We found out that in the group defined by low energy and identified by conformational selection mechanism, the majority of the contacts are with arginine and the DNA phosphate (71%), while only 24% of the protein

interacts with the bases. This protein-DNA binding is mainly driven by the electrostatics and the shape of the DNA. In the induced fit group, 34% of the interactions involve the bases, while only 51% involves the phosphates; suggesting that the protein changes the free DNA conformation to increase the interaction between the protein and the bases. This is confirmed by the increase of amidic residues present at the interface (GLN and ASN), which are very well suited to form direct bonds with the DNA bases (Supplementary Fig. S10 and scheme of recognition modes in Supplementary Fig. S11). Furthermore our atomistic and mesoscopic analysis suggest that in our cases DNA-interacting proteins could have evolved to recognize the native shape of the DNA duplex, avoiding the need to invest large amounts of energy in deforming the native physiological B-DNA, which would make the effector protein less efficient when competing with histones, RNAs, and many other proteins.

To evaluate the generality of our conclusions we hand-curate several (17) complexes that were excluded from the initial analysis as the PDB dataset contained unpaired or modified bases (see Supplementary Methods). Supplementary Fig. S12 shows that also for these complexes the energy values fall within the range expected by the conformational selection paradigm (green bars).

What is the role of base or shape readout in protein binding to cognate DNA sequences? To answer this fundamental question we compared the distribution of deformation energies of one million randomly generated sequences with that of the DNA sequences of our set of 50 representative complexes. Results in Fig. 7a show that the energetic cost for reaching the bioactive state for DNA sequences found in PDB is lower than for random sequences. So, it appears for these cases that the sequences in

the Xray crystal complex can reach the bioactive (bound) state much more easily than random sequences. As DNA sequences used to solve structures deposited in PDB structures tend to be consensus sequences, we can guess that, in general, the shape readout model plays a major role in selecting cognate sequences. To further validate this hypothesis, we repeated the study considering a further 20 sequences fulfilling the consensus sequence requirements taken from *in vitro* footprinting/SELEX experiments (http://floresta.eead.csic.es/footprintdb/)[51]. Results show again that the positioning of the sequences with consensus pattern, those sequences position at lower energy costs with respect to the random ones and are largely favoured (green lines in Fig. 7b) energetically (grey in Fig. 7b) to achieve the bound conformation. This confirms that *in vitr*o high affinity sequences are typically those showing less resistance to be distorted by the protein, as expected by the shape readout binding mechanism. In summary, theoretical results strongly favour the shape reading mechanism as a major contributor to the selection of DNA binding sequences and that nucleotide sequence alone does not fully explain the widely observed mechanism of DNA shape readout..

Materials and Methods

DNA-protein complex selection. The dataset representing the DNA-protein complexes was obtained after applying a set of filters to the whole collection of protein complexes deposited in the Protein Data Bank (PDB, www.rcsb.org) [52]. The initial dataset was acquired from Nucleic Acid Database (NDB) [53], selecting PDB entries having protein molecules attached to double-stranded B-DNA, thus avoiding single-stranded nucleic acid structures, RNA, and non-canonical B-DNA conformations. From this initial set we removed protein redundancy and selected the 1,038 unique protein-B-DNA entries found

[52]. We then filtered this set excluding DNAs with modified nucleic bases, unpairing or mismatches, broken strands or non-Watson-Crick pairing (details about PDB filtering in Supplementary Material, Supplementary Methods and Supplementary Fig. S1), obtaining an dataset of 174 protein-DNA complexes that in this work defines the protein-DNA interactome. The interactions involved in the structures of this dataset have been further studied using the R package VeriNA3D [54] in Supplementary Fig. S11. From this set, we selected a sample of 50 diverse cases from the PDB (see Fig. 1 and Supplementary Table S1 for details[7]) covering different types of protein folds and function, DNA recognition modes (minor/major grooves), sequence binding motifs and structural selection. We extracted the DNA sequences from the selected 50 protein-DNA complexes and those sequences were subjected to atomistic MD simulations in *in silico* physiological conditions.

Atomistic Simulations. Starting models for all the protein-free DNAs were created using Arnott-B DNA canonical values [55]. Additionally, as a way to control convergence, for a few systems trajectories were also started from the DNA conformation in the protein-DNA complex. Each system was solvated using TIP3P waters [56] in a truncated octahedron box with periodic boundary conditions, and adding Na^+ ions until neutralization and extra salt, up to 0.15 M in NaCl, using Smith and Dang ion parameters [57]. The DNA interactions were represented using parmbsc1 force-field [40–42]. All simulations were performed using Amber 14 suite of programs (AMBER 2014 San Francisco University of California). The systems were then energy minimized, thermalized and pre-equilibrated using our standard multi-step protocol [42,58] followed by 50 ns of equilibration before 0.5 μs of unbiased MD simulations using standard simulation conditions in the NPT ensemble (see Supplementary

Methods). Trajectories and associated analysis are deposited in the MuG-BigNASim database [59] (http://mmb.irbbarcelona.org/BigNASim/) and are freely accessible.

Analysis of trajectories. Collected trajectories (500,000 structures per system) were post-processed and analysed using the CPPTRAJ module of the Ambertools package [60], the NAFlex server [61], VMD 1.9, Bio3D R library [62], PCAsuite [63] and Curves+ package [64], as well as "in house" software. The interaction potential (electrostatics and van der Waals) of Na^+ and $Na^+(H_2O)_6$ probes with DNA duplexes was determined using a linear approximation to the Poisson-Boltzmann equation and dielectric constant for the DNA $\varepsilon_{DNA} = 8$ [65], as implemented in the CMIP program [66].

Statistical analysis of base pairs parameters: Hotelling´s multivariate statistical test [67] was used to analyse whether or not the distribution of a given helical parameter in the DNA-protein complex fits the expected distribution in the naked-DNA conformational ensemble. Accordingly, multivariate F statistic was defined [67] as:

$$F = \frac{n-m}{m}(\mu - \bar{x})^t S^{-1}(\mu - \bar{x}) \qquad \text{(eq. 1)}$$

where μ is the vector containing the m experimental values for each base pair step along the sequence taken from the complex structure. From the matrix (n×m) containing the values for each base pair step parameter (m) obtained from the n time-frames of the MD simulations, the average values along the time ($\bar{x}$) and the inverse of the variance matrix (S^{-1}) have been calculated. Following Hotelling statistical test, the bound conformation is considered not sampled by the naked DNA trajectory when the computed F falls outside the confidence region (CR) $F > F_{(1-\alpha;m,n-m)}$ at 1-α = 95 % confidence level,

where $F_{(1-\alpha;m,n-m)}$ is the quantile 1-α from an F distribution with m, n-m degrees of freedom.

Essential dynamics analysis: Essential dynamics (ED) [68,69] analysis has been performed to determine the essential movements explaining the DNA global dynamics [68,69]. Eigenvectors ($\overrightarrow{\vartheta_i}$ in eq. 2) and eigenvalues were determined by diagonalization of the covariance matrix following the R package Bio3D [62]. The reduced set of eigenvectors that explain 90% of the variance (n in equation 2), have been selected in each case as descriptive of the essential dynamics of the duplexes. The ability of the essential dynamics of DNA to trace the conformational transition from the unbound to the bound state (given by vector $\vec{R}$ in eq. 2) was measured by the cumulative sum of the squared overlap (γ) between the transition vector and the eigenvectors describing the essential dynamics of the naked duplex [70,71]:

$$\gamma = \sum_{i=1}^{n}\left(\overrightarrow{\vartheta_i} \cdot \vec{R}\right)^2 \qquad \text{(eq. 2)}$$

An additional measure of the ability of the essential deformation space of DNA to reproduce a transition is given by the percentage of the transition (Distance covered measured using the RMSD) that can be achieved by moving along the -*n*- essential deformation modes. In other words, how close to the bioactive conformation can the DNA arrive when moving across the easiest deformation modes (equation 3):

$$Distance\ covered = \frac{RMSD_{in} - RMSD_{90\%}}{RMSD_{in}}\% \qquad \text{(eq. 3)}$$

where $RMSD_{in}$ is calculated between the naked DNA and the protein bound conformations. $RMSD_{fin}$ is the minimum RMSD between the bound structure and the naked DNA after the displacement along the essential deformation modes that describe 90% of the naked DNA motion.

Deformation energy analysis: The deformation energy associated to the DNA transition from naked to bound is approximated in the harmonic regime [72]:

$$Def.Energy = \frac{\sum_{j=1}^{m} E_j}{m}, \text{ with } E_j = \frac{1}{2}\sum_{s=1}^{6}\sum_{t=1}^{6} k_{st}^{j}\Delta X_s^j\,\Delta X_t^j \qquad \text{(eq. 4)}$$

where j stands for each of the m base pair steps of the DNA. In turn, E_j is determined from a stiffness mesoscopic model [72–74], where ΔX_s^j and ΔX_t^j are the deviation from equilibrium values in the 6 base pair step helical parameters (roll, twist, tilt, slide, rise or shift) and k_{st}^j stands for the elements of the stiffness matrix obtained by inversion of the MD covariance matrix in the helical space, as determined by Olson-Lankaš model [73–75]. The equilibrium values and stiffness constants for each individual base pair step [40,75,76] were taken from a MD simulations stored in the BigNASim [59] that cover all the unique base pair steps in all the possible tetranucleotide environments from microsecond-long parmbsc1 simulations. Estimates of deformation energy associated to the change from the naked to the bound conformation where compared with the thermal energy fluctuation of naked B-DNA in solution. The thermal energy fluctuation of naked B-DNA is the deformation energy sampled by the DNA at room temperature. For each snapshot of each free DNA trajectory we computed the deformation respect to the corresponding average structure. The thermal energy

fluctuation is then defined taking the average plus one standard deviation from the distribution built from the collection of these energy values. The thermal energy fluctuation determined in this study from all the simulated systems amounts to 2.5 ± 0.1 kcal/mol·bp.

Acknowledgments

M.O. is an ICREA (Institució Catalana de Recerca i Estudis Avancats) academia researcher. P.D.D. is a PEDECIBA (Programa de Desarrollo de las Ciencias Basicas) and SNI (Sistema Nacional de Investigadores, Agencia Nacional de Investigación e Innovación, Uruguay) researcher. This work was supported by the Spanish Ministry of Science [grants BIO2015-64802-R-]; Spanish Ministry of Science (BFU2014-61670-EXP and BFU2014-52864-R); the Catalan Government [grants 2014-SGR]; the Instituto de Salud Carlos III-Instituto Nacional de Bioinformática; the European Union's Horizon 2020 research and innovation program (676556), and the Biomolecular and Bioinformatics Resources Platform (ISCIII PT 13/000/0030 co-funded by the Fondo Europeo de Desarrollo Regional [FEDER]) [grants Elixir-Excelerate: 676559; BioExcel2:823830 and MuG: 676566]. Funding was also provided by the MINECO Severo Ochoa Award of Excel- lence from the Government of Spain (awarded to IRB Barcelona).

References

[1] R. Rohs, X. Jin, S.M. West, R. Joshi, B. Honig, R.S. Mann, Origins of specificity in protein-DNA recognition., Annu. Rev. Biochem. 79 (2010) 233–69.

doi:10.1146/annurev-biochem-060408-091030.

[2] A. Tóth-Petróczy, I. Simon, M. Fuxreiter, Y. Levy, Disordered Tails of Homeodomains Facilitate DNA Recognition by Providing a Trade-Off between Folding and Specific Binding, J. Am. Chem. Soc. 131 (2009) 15084–15085. doi:10.1021/ja9052784.

[3] L.-A. Harris, L.D. Williams, G.B. Koudelka, Specific minor groove solvation is a crucial determinant of DNA binding site recognition., Nucleic Acids Res. 42 (2014) 14053–9. doi:10.1093/nar/gku1259.

[4] A. Debnath, B. Mukherjee, K.G. Ayappa, P.K. Maiti, S.-T. Lin, Entropy and dynamics of water in hydration layers of a bilayer, J. Chem. Phys. 133 (2010) 174704. doi:10.1063/1.3494115.

[5] G.G. Hammes, Y.-C. Chang, T.G. Oas, Conformational selection or induced fit: a flux description of reaction mechanism., Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 13737–41. doi:10.1073/pnas.0907195106.

[6] H.-X. Zhou, From induced fit to conformational selection: a continuum of binding mechanism controlled by the timescale of conformational transitions., Biophys. J. 98 (2010) L15-7. doi:10.1016/j.bpj.2009.11.029.

[7] J. Li, J.M. Sagendorf, T.-P. Chiu, M. Pasi, A. Perez, R. Rohs, Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding., Nucleic Acids Res. 45 (2017) 12877–12887. doi:10.1093/nar/gkx1145.

[8] M.F. Berger, M.L. Bulyk, Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors, Nat. Protoc. 4 (2009) 393–411. doi:10.1038/nprot.2008.195.

[9] C. Zhu, K.J.R.P. Byers, R.P. McCord, Z. Shi, M.F. Berger, D.E. Newburger, K. Saulrieta, Z. Smith, M. V Shah, M. Radhakrishnan, A.A. Philippakis, Y. Hu, F. De Masi, M.

Pacek, A. Rolfs, T. Murthy, J. Labaer, M.L. Bulyk, E. Fraenkel, R. Young, High-resolution DNA-binding specificity analysis of yeast transcription factors., Genome Res. 19 (2009) 556–66. doi:10.1101/gr.090233.108.

[10] G. Badis, E.T. Chan, H. van Bakel, L. Pena-Castillo, D. Tillo, K. Tsui, C.D. Carlson, A.J. Gossett, M.J. Hasinoff, C.L. Warren, M. Gebbia, S. Talukder, A. Yang, S. Mnaimneh, D. Terterov, D. Coburn, A. Li Yeo, Z.X. Yeo, N.D. Clarke, J.D. Lieb, A.Z. Ansari, C. Nislow, T.R. Hughes, A Library of Yeast Transcription Factor Motifs Reveals a Widespread Function for Rsc3 in Targeting Nucleosome Exclusion at Promoters, Mol. Cell. 32 (2008) 878–887. doi:10.1016/j.molcel.2008.11.020.

[11] M.B. Noyes, R.G. Christensen, A. Wakabayashi, G.D. Stormo, M.H. Brodsky, S.A. Wolfe, Analysis of Homeodomain Specificities Allows the Family-wide Prediction of Preferred Recognition Sites, Cell. 133 (2008) 1277–1289. doi:10.1016/j.cell.2008.05.023.

[12] A. Arvey, P. Agius, W.S. Noble, C. Leslie, Sequence and chromatin determinants of cell-type-specific transcription factor binding., Genome Res. 22 (2012) 1723–34. doi:10.1101/gr.127712.111.

[13] D.E. Newburger, M.L. Bulyk, UniPROBE: an online database of protein binding microarray data on protein-DNA interactions, Nucleic Acids Res. 37 (2009) D77–D82. doi:10.1093/nar/gkn660.

[14] Z. Xie, S. Hu, S. Blackshaw, H. Zhu, J. Qian, hPDI: a database of experimental human protein-DNA interactions, Bioinformatics. 26 (2010) 287–289. doi:10.1093/bioinformatics/btp631.

[15] S. Mahony, B.F. Pugh, Protein–DNA binding in high-resolution, Crit. Rev. Biochem. Mol. Biol. 50 (2015) 269–283. doi:10.3109/10409238.2015.1051505.

[16] J. Li, J.M. Sagendorf, T.-P. Chiu, M. Pasi, A. Perez, R. Rohs, Expanding the repertoire

of DNA shape features for genome-scale studies of transcription factor binding., Nucleic Acids Res. 45 (2017) 12877–12887. doi:10.1093/nar/gkx1145.

[17] T.R. Riley, M. Slattery, N. Abe, C. Rastogi, D. Liu, R.S. Mann, H.J. Bussemaker, SELEX-seq: A Method for Characterizing the Complete Repertoire of Binding Site Preferences for Transcription Factor Complexes, in: Methods Mol. Biol., 2014: pp. 255–278. doi:10.1007/978-1-4939-1242-1_16.

[18] A. Jolma, J. Yan, T. Whitington, J. Toivonen, K.R. Nitta, P. Rastas, E. Morgunova, M. Enge, M. Taipale, G. Wei, K. Palin, J.M. Vaquerizas, R. Vincentelli, N.M. Luscombe, T.R. Hughes, P. Lemaire, E. Ukkonen, T. Kivioja, J. Taipale, DNA-Binding Specificities of Human Transcription Factors, Cell. 152 (2013) 327–339. doi:10.1016/j.cell.2012.12.009.

[19] V.A. Kuznetsov, O. Singh, P. Jenjaroenpun, Statistics of protein-DNA binding and the total number of binding sites for a transcription factor in the mammalian genome, BMC Genomics. 11 (2010) S12. doi:10.1186/1471-2164-11-S1-S12.

[20] M. Yu, G.C. Hon, K.E. Szulwach, C.-X. Song, L. Zhang, A. Kim, X. Li, Q. Dai, Y. Shen, B. Park, J.-H. Min, P. Jin, B. Ren, C. He, Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome., Cell. 149 (2012) 1368–80. doi:10.1016/j.cell.2012.04.027.

[21] Z. Miao, E. Westhof, Prediction of nucleic acid binding probability in proteins: a neighboring residue network based score, Nucleic Acids Res. 43 (2015) 5340–5351. doi:10.1093/nar/gkv446.

[22] A. van der Vaart, Coupled binding–bending–folding: The complex conformational dynamics of protein-DNA binding studied by atomistic molecular dynamics simulations, Biochim. Biophys. Acta - Gen. Subj. 1850 (2015) 1091–1098. doi:10.1016/J.BBAGEN.2014.08.009.

[23] K.M. Thayer, D.L. Beveridge, Hidden Markov models from molecular dynamics simulations on DNA, Proc. Natl. Acad. Sci. 99 (2002) 8642–8647. doi:10.1073/pnas.132148699.

[24] Y. Zhao, D. Granas, G.D. Stormo, D. Johnson, R. Myers, Inferring Binding Energies from Selected Binding Sites, PLoS Comput. Biol. 5 (2009) e1000590. doi:10.1371/journal.pcbi.1000590.

[25] D.J. Galas, A. Schmitz, DNAse footprinting: a simple method for the detection of protein-DNA binding specificity., Nucleic Acids Res. 5 (1978) 3157–70. http://www.ncbi.nlm.nih.gov/pubmed/212715 (accessed July 24, 2017).

[26] R.B. Jones, A. Gordus, J.A. Krall, G. MacBeath, A quantitative protein interaction network for the ErbB receptors using protein microarrays, Nature. 439 (2006) 168–174. doi:10.1038/nature04177.

[27] C.T. Harbison, D.B. Gordon, T.I. Lee, N.J. Rinaldi, K.D. Macisaac, T.W. Danford, N.M. Hannett, J.-B. Tagne, D.B. Reynolds, J. Yoo, E.G. Jennings, J. Zeitlinger, D.K. Pokholok, M. Kellis, P.A. Rolfe, K.T. Takusagawa, E.S. Lander, D.K. Gifford, E. Fraenkel, R.A. Young, Transcriptional regulatory code of a eukaryotic genome, Nature. 431 (2004) 99–104. doi:10.1038/nature02800.

[28] D.S. Johnson, A. Mortazavi, R.M. Myers, B. Wold, Genome-Wide Mapping of in Vivo Protein-DNA Interactions, Science (80-.). 316 (2007).

[29] R. Rohs, S.M. West, A. Sosinsky, P. Liu, R.S. Mann, B. Honig, The role of DNA shape in protein-DNA recognition., Nature. 461 (2009) 1248–53. doi:10.1038/nature08473.

[30] G. Paillard, R. Lavery, Analyzing Protein-DNA Recognition Mechanisms, Structure. 12 (2004) 113–122. doi:10.1016/j.str.2003.11.022.

[31] C. Chen, B.M. Pettitt, The binding process of a nonspecific enzyme with DNA.,

Biophys. J. 101 (2011) 1139–47. doi:10.1016/j.bpj.2011.07.016.

[32] A.N. Temiz, P. V Benos, C.J. Camacho, Electrostatic hot spot on DNA-binding domains mediates phosphate desolvation and the pre-organization of specificity determinant side chains., Nucleic Acids Res. 38 (2010) 2134–44. doi:10.1093/nar/gkp1132.

[33] B. Bouvier, K. Zakrzewska, R. Lavery, Protein-DNA Recognition Triggered by a DNA Conformational Switch, Angew. Chemie Int. Ed. 50 (2011) 6516–6518. doi:10.1002/anie.201101417.

[34] S. Furini, P. Barbini, C. Domene, DNA-recognition process described by MD simulations of the lactose repressor protein on a specific and a non-specific DNA sequence, Nucleic Acids Res. 41 (2013) 3963–3972. doi:10.1093/nar/gkt099.

[35] I. Sela, D.B. Lukatsky, C. Nislow, et al., A.M. van Oijen, et al., DNA sequence correlations shape nonspecific transcription factor-DNA binding affinity., Biophys. J. 101 (2011) 160–6. doi:10.1016/j.bpj.2011.04.037.

[36] M. Fuxreiter, I. Simon, S. Bondos, Dynamic protein–DNA recognition: beyond what can be seen, Trends Biochem. Sci. 36 (2011) 415–423. doi:10.1016/j.tibs.2011.04.006.

[37] L. Etheve, J. Martin, R. Lavery, Dynamics and recognition within a protein–DNA complex: a molecular dynamics study of the SKN-1/DNA interaction, Nucleic Acids Res. 44 (2016) 1440–1448. doi:10.1093/nar/gkv1511.

[38] D.D. Boehr, R. Nussinov, P.E. Wright, The role of dynamic conformational ensembles in biomolecular recognition, Nat. Chem. Biol. 5 (2009) 789–796. doi:10.1038/nchembio.232.

[39] T.-P. Chiu, F. Comoglio, T. Zhou, L. Yang, R. Paro, R. Rohs, DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding, (n.d.).

doi:10.1093/bioinformatics/btv735.

[40] I. Ivani, P.D. Dans, A. Noy, A. Pérez, I. Faustino, A. Hospital, J. Walther, P. Andrio, R. Goñi, A. Balaceanu, G. Portella, F. Battistini, J.L. Gelpí, C. González, M. Vendruscolo, C.A. Laughton, S.A. Harris, D.A. Case, M. Orozco, Parmbsc1: a refined force field for DNA simulations, Nat. Methods. 13 (2015) 55–8. doi:10.1038/nmeth.3658.

[41] P.D. Dans, I. Ivani, A. Hospital, G. Portella, C. González, M. Orozco, How accurate are accurate force-fields for B-DNA?, Nucleic Acids Res. 45 (2017) gkw1355. doi:10.1093/nar/gkw1355.

[42] P.D. Dans, L. Danilāne, I. Ivani, T. Dršata, F. Lankaš, A. Hospital, J. Walther, R.I. Pujagut, F. Battistini, J.L. Gelpí, R. Lavery, M. Orozco, Long-timescale dynamics of the Drew–Dickerson dodecamer, Nucleic Acids Res. 44 (2016) 4052–4066. doi:10.1093/nar/gkw264.

[43] A. Kuzmanic, P.D. Dans, M. Orozco, An In-Depth Look at DNA Crystals through the Prism of Molecular Dynamics Simulations, Chem. (2019). doi:10.1016/J.CHEMPR.2018.12.007.

[44] P.D. Dans, J. Walther, H. Gómez, Multiscale simulation of DNA, Curr. Opin. Struct. Biol. 37 (2016) 29–45. doi:10.1016/J.SBI.2015.11.011.

[45] L.J.W. Murray, W.B. Arendall, D.C. Richardson, J.S. Richardson, RNA backbone is rotameric., Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 13904–9. doi:10.1073/pnas.1835769100.

[46] A. Ben Imeddourene, X. Xu, L. Zargarian, C. Oguey, N. Foloppe, O. Mauffret, B. Hartmann, The intrinsic mechanics of B-DNA in solution characterized by NMR., Nucleic Acids Res. 44 (2016) 3432–47. doi:10.1093/nar/gkw084.

[47] R. Galindo-Murillo, J.C. Robertson, M. Zgarbová, J. Šponer, M. Otyepka, P. Jurečka, T.E. Cheatham, Assessing the Current State of Amber Force Field Modifications for

DNA, J. Chem. Theory Comput. 12 (2016) 4114–4127. doi:10.1021/acs.jctc.6b00186.

[48] P.D. Dans, A. Pérez, I. Faustino, R. Lavery, M. Orozco, Exploring polymorphisms in B-DNA helical conformations., Nucleic Acids Res. 40 (2012) 10668–78. doi:10.1093/nar/gks884.

[49] P.D. Dans, I. Faustino, F. Battistini, K. Zakrzewska, R. Lavery, M. Orozco, Unraveling the sequence-dependent polymorphic behavior of d (CpG) steps in B-DNA, Nucleic Acids Res. 42 (2015) 11304–11320.

[50] N. Abe, I. Dror, L. Yang, M. Slattery, T. Zhou, H.J. Bussemaker, R. Rohs, R.S. Mann, Deconvolving the Recognition of DNA Shape from Sequence, Cell. 161 (2015) 307–318. doi:10.1016/J.CELL.2015.02.008.

[51] B. Contreras-Moreira, 3D-footprint: a database for the structural analysis of protein-DNA complexes., Nucleic Acids Res. 38 (2010) D91-7. doi:10.1093/nar/gkp781.

[52] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank, Nucleic Acids Res. 28 (2000) 235–242. doi:10.1093/nar/28.1.235.

[53] B. Coimbatore Narayanan, J. Westbrook, S. Ghosh, A.I. Petrov, B. Sweeney, C.L. Zirbel, N.B. Leontis, H.M. Berman, The Nucleic Acid Database: new features and capabilities, Nucleic Acids Res. 42 (2014) D114–D122. doi:10.1093/nar/gkt980.

[54] D. Gallego, L. Darré, P.D. Dans, M. Orozco, VeriNA3d: An R Package for Nucleic Acids Data Mining, Bioinformatics. (2019). doi:10.1093/bioinformatics/btz553.

[55] S. Arnott, D.W. Hukins, Optimised parameters for A-DNA and B-DNA., Biochem. Biophys. Res. Commun. 47 (1972) 1504–9. http://www.ncbi.nlm.nih.gov/pubmed/5040245 (accessed May 22, 2017).

[56] W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey, M.L. Klein, Comparison of simple potential functions for simulating liquid water, J. Chem. Phys. 79 (1983) 926–935. doi:10.1063/1.445869.

[57] D.E. Smith, L.X. Dang, Computer simulations of NaCl association in polarizable water, J. Chem. Phys. 100 (1994) 3757–3766. doi:10.1063/1.466363.

[58] A. Pérez, F.J. Luque, M. Orozco, Dynamics of B-DNA on the Microsecond Time Scale, J. Am. Chem. Soc. 129 (2007) 14739–14745. doi:10.1021/ja0753546.

[59] A. Hospital, P. Andrio, C. Cugnasco, L. Codo, Y. Becerra, P.D. Dans, F. Battistini, J. Torres, R. Goñi, M. Orozco, J.L. Gelpí, BIGNASim: a NoSQL database structure and analysis portal for nucleic acids simulation data, Nucleic Acids Res. 44 (2016) D272–D278. doi:10.1093/nar/gkv1301.

[60] D.R. Roe, T.E. Cheatham, PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data, J. Chem. Theory Comput. 9 (2013) 3084–3095. doi:10.1021/ct400341p.

[61] A. Hospital, I. Faustino, R. Collepardo-Guevara, C. Gonzalez, J.L. Gelpi, M. Orozco, NAFlex: a web server for the study of nucleic acid flexibility, Nucleic Acids Res. 41 (2013) W47–W55. doi:10.1093/nar/gkt378.

[62] L. Skjærven, X.-Q. Yao, G. Scarabelli, B.J. Grant, Integrating protein structural dynamics and evolutionary analysis with Bio3D, BMC Bioinformatics. 15 (2014) 399. doi:10.1186/s12859-014-0399-6.

[63] T. Meyer, C. Ferrer-Costa, A. Pérez, M. Rueda, A. Bidon-Chanal, F.J. Luque, A. Charles. A. Laughton, M. Orozco, Essential Dynamics: A Tool for Efficient Trajectory Compression and Management, (2006). doi:10.1021/CT050285B.

[64] R. Lavery, M. Moakher, J.H. Maddocks, D. Petkeviciute, K. Zakrzewska, Conformational analysis of nucleic acids revisited: Curves+., Nucleic Acids Res. 37

(2009) 5917–29. doi:10.1093/nar/gkp608.

[65] A. Cuervo, P.D. Dans, J.L. Carrascosa, M. Orozco, G. Gomila, L. Fumagalli, Direct measurement of the dielectric polarization properties of DNA, Proc. Natl. Acad. Sci. 111 (2014) E3624–E3630. doi:10.1073/pnas.1405702111.

[66] J.L. Gelpí, S.G. Kalko, X. Barril, J. Cirera, X. de La Cruz, F.J. Luque, M. Orozco, Classical molecular interaction potentials: improved setup procedure in molecular dynamics simulations of proteins., Proteins. 45 (2001) 428–37. http://www.ncbi.nlm.nih.gov/pubmed/11746690 (accessed May 24, 2017).

[67] W. Härdle, L. Simar, Applied multivariate statistical analysis, n.d.

[68] M. Orozco, A. Pérez, A. Noy, F.J. Luque, Theoretical methods for the simulation of nucleic acids, Chem. Soc. Rev. 32 (2003) 350–364. doi:10.1039/B207226M.

[69] A. Amadei, A.B.M. Linssen, H.J.C. Berendsen, Essential dynamics of proteins, Proteins Struct. Funct. Genet. 17 (1993) 412–425. doi:10.1002/prot.340170408.

[70] A. Pérez, J.R. Blas, M. Rueda, J.M. López-Bes, X. de la Cruz, M. Orozco, Exploring the Essential Dynamics of B-DNA., J. Chem. Theory Comput. 1 (2005) 790–800. doi:10.1021/ct050051s.

[71] A. Noy, F. Javier Luque, M. Orozco, Theoretical Analysis of Antisense Duplexes: Determinants of the RNase H Susceptibility, (2008). doi:10.1021/JA076734U.

[72] G. Portella, F. Battistini, M. Orozco, Understanding the connection between epigenetic DNA methylation and nucleosome positioning from computer simulations., PLoS Comput. Biol. 9 (2013) e1003354. doi:10.1371/journal.pcbi.1003354.

[73] F. Lankas, J. Sponer, J. Langowski, T.E. Cheatham, DNA basepair step deformability inferred from molecular dynamics simulations., Biophys. J. 85 (2003) 2872–83. doi:10.1016/S0006-3495(03)74710-9.

[74] A. Pérez, F. Lankas, F.J. Luque, M. Orozco, Towards a molecular dynamics consensus view of B-DNA flexibility., Nucleic Acids Res. 36 (2008) 2379–94. doi:10.1093/nar/gkn082.

[75] W.K. Olson, DNA sequence-dependent deformability deduced from protein-DNA crystal complexes, Proc. Natl. Acad. Sci. 95 (1998) 11163–11168. doi:10.1073/pnas.95.19.11163.

[76] F. Lankas, J. Sponer, J. Langowski, T.E. Cheatham 3rd, DNA basepair step deformability inferred from molecular dynamics simulations, Biophys. J. 85 (2003) 2872–2883.

Figures Legends

Fig. 1. Representation of the protein-DNA complexes summarised with details in Table S1[7]. The PDB IDs are indicated.

Fig. 2. Base pair parameter confidence region profile. For each protein-bound DNA structure identified by their PDB ID, the axis represents the difference between the observed test statistic and the 95% critical value from the F distribution ($F - F_{(1-\alpha;m,n-m)}$). The value for each base pair parameter, translation rise and rotational roll, can be inside (<0, limit defined by red line) or outside the naked DNA conformational space (>0). See Methods and Supp. Methods for discussion.

Fig. 3. Backbone and base pair parameter analysis for the complex PDB ID 1A0A. (a) Analysis of the backbone angles (α,β,χ,ε,γ,phase, ζ) is shown. Backbone angles variation has been analysed using the difference between the experimental protein-bound DNA angle values and the average MD simulated naked DNA values plus the standard deviation, divided by the standard deviation along the MD trajectory for each backbone angle (Δ(bound-MDnaked)). (b) comparison between the experimental (blue) and MD values (red with standard deviation contour in pink) for base pair step parameter roll and rise. The distortion given by the contact of the protein helices and coil residues (in red and blue respectively in the image on top and named in the left panel) at the base level, extreme roll and rise values at steps CC, CG and GT, is correlated with deformation of the backbone angles in the backbone.

Fig. 4. Correlation between the $RMSD_{in}$, calculated between the average conformation along the MD simulation of the unbound DNA and experimental protein-bound structure of the DNA for the complexes studied (Table S1), and: (a) Deformation energy cost (kcal/mol·bp) to move from the unbound to the experimental bound (bioactive) conformation in the helical space; (b) Overlap squared between the essential dynamics of the unbound DNA and vector that connects the unbound and bound conformations; (c) Distance covered when moving the unbound structure along the essential modes (those describing 90% of naked simulation variance) towards the bound (bioactive) structure; (d) RMSD with bound (bioactive) conformation after moving the naked structure along the essential modes in the direction of the bound (bioactive) conformation ($RMSD_{fin}$). The bound DNA structures detected with probable

uncertainties in the experimental structure are highlighted in red, the protein DNA-complexes with DNA distorted by the protein in yellow and the remaining systems are represented with blue dots. We marked with PDB ID names the structure with deformation energy higher than 5 kcal/mol·bp.

Fig. 5. Molecular Interaction Potential (MIP) using Na^+ as probe for 3 cases, the unbound (left column, upper image) and bound (left column bottom image and right column) DNA structures. The isosurfaces (in red) have been calculated for the DNA sequences in the complexes that showed the mostly distorted structures (right column) in our dataset: (a) 1J46 (isovalue =-6.4 kcal mol^{-1}), (b) 3F27 (isovalue =-7.4 kcal mol^{-1}), (c) 1CDW (isovalue =-7.4 kcal mol^{-1}). In the right column details of the protein residues with positive charges (lysines and arginines in licorice) pointing in the direction of the detected potential surface are represented.

Fig. 6. Frequency of the deformation energy cost (kcal/mol·bp) required moving from the unbound to the bound conformation in the helical space for all the DNA-protein interactome. In red images of structures that require high Deformation Energy: bent and very distorted structures at the backbone and base pair step level (high roll value), 1YA6, 2ADW, 5H1C PDB ID respectively from lower to higher energy. In this distribution the number of structures that fall within the area with energy < 2.5 kcal/mol·bp (blue area) and energy between 2.5 and 5 kcal/mol·bp (red area) are represented. Percentage for the whole selected interactome is shown in the right top corner.

Fig. 7. (a) Relative position of the deformation energy value of the essayed PDB sequences in the frequency distribution of energies for a million random sequences (see Supplementary Methods). Highlighted with corresponding letters (A-F) some examples, with the respective distributions on the right, identified with PDB code: in grey the energy distribution for the random sequences and in yellow the sequence found in the experimental complex. (b) Comparison between the distribution of the deformation energies for random sequences (grey) and the deformation energy sequences with experimental high-affinity pattern (green) as found in Footprint Database.

Highlights

- Deciphering protein-DNA recognition using DNA dynamics.
- Conformational selection and induced fit differentiation in some protein-DNA complexes.
- Direct vs Indirect recognition mode for some protein data bank protein-DNA complexes.
- Consensus sequences selection based on structure and energetics.

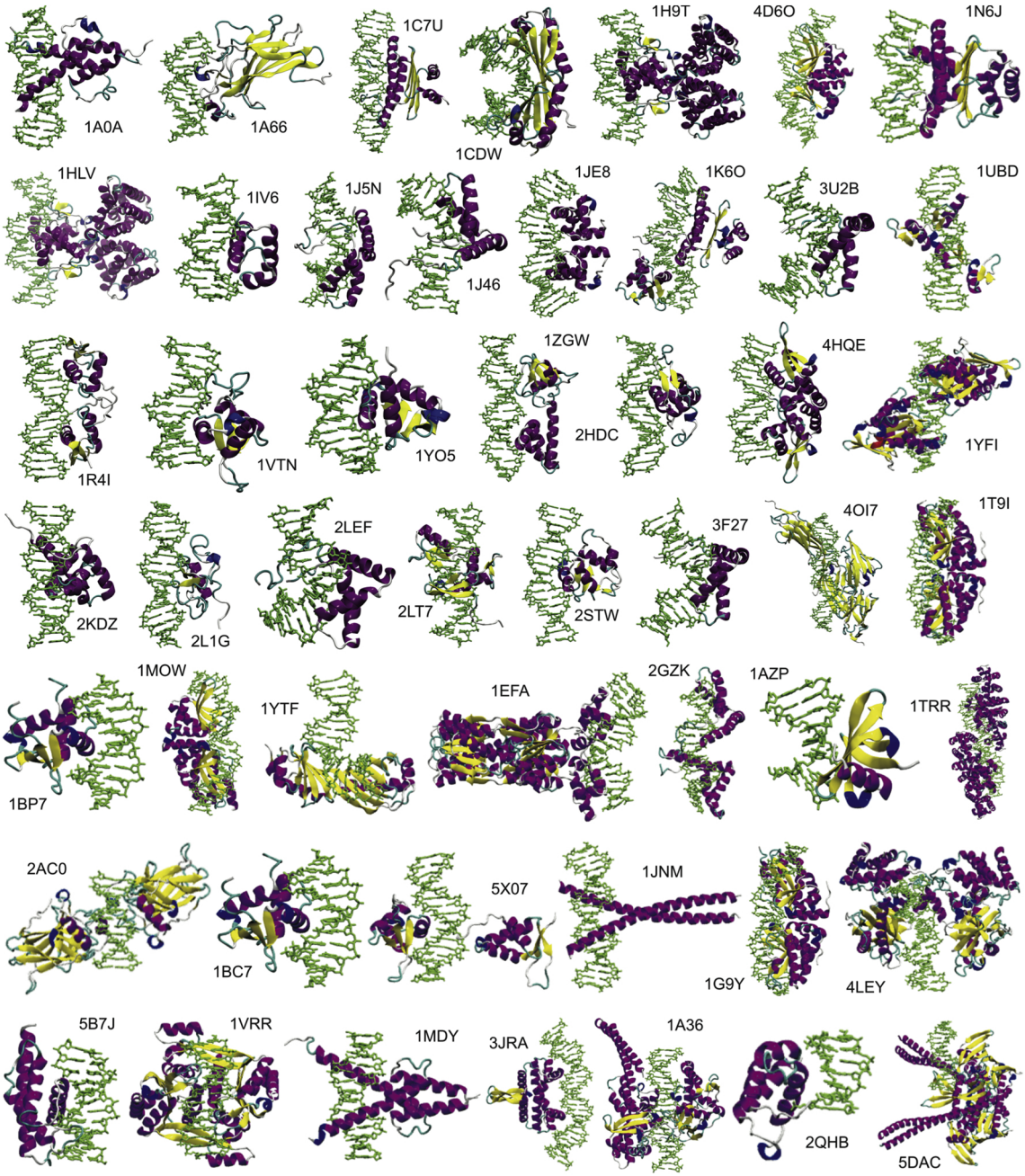
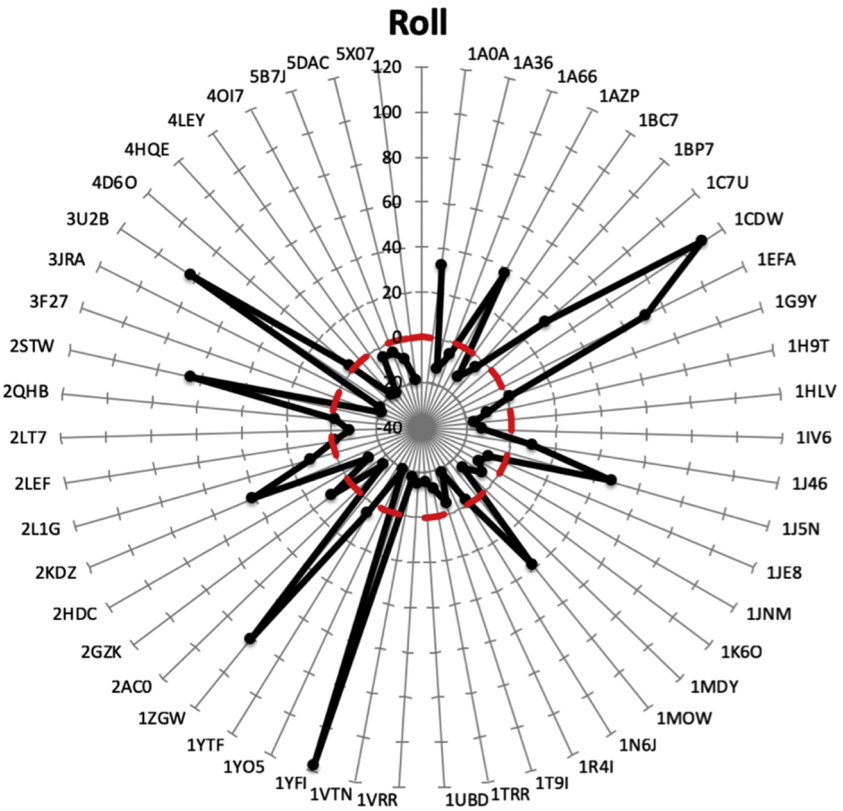

1A0A
1A66
1C7U
1CDW
1H9T
4D6O
1N6J
1HLV
1IV6
1J5N
1J46
1JE8
1K6O
3U2B
1UBD
1R4I
1VTN
1YO5
1ZGW
2HDC
4HQE
1YFI
2KDZ
2L1G
2LEF
2LT7
2STW
3F27
4OI7
1T9I
1BP7
1MOW
1YTF
1EFA
2GZK
1AZP
1TRR
2AC0
1BC7
5X07
1JNM
1G9Y
4LEY
5B7J
1VRR
1MDY
3JRA
1A36
2QHB
5DAC

Figure 1


Roll
5X07
1A0A
1A36
1A66
1AZP
1BC7
1BP7
1C7U
1CDW
1EFA
1G9Y
1H9T
1HLV
1IV6
1J46
1J5N
1JE8
1JNM
1K6O
1MDY
1MOW
1N6J
1R4I
1T9I
1TRR
1UBD
1VRR
1VTN
1YFI
1YO5
1YTF
1ZGW
2AC0
2GZK
2HDC
2KDZ
2L1G
2LEF
2LT7
2QHB
2STW
3F27
3JRA
3U2B
4D6O
4HQE
4LEY
4OI7
5B7J
5DAC
120
100
80
60
40
20
0
40

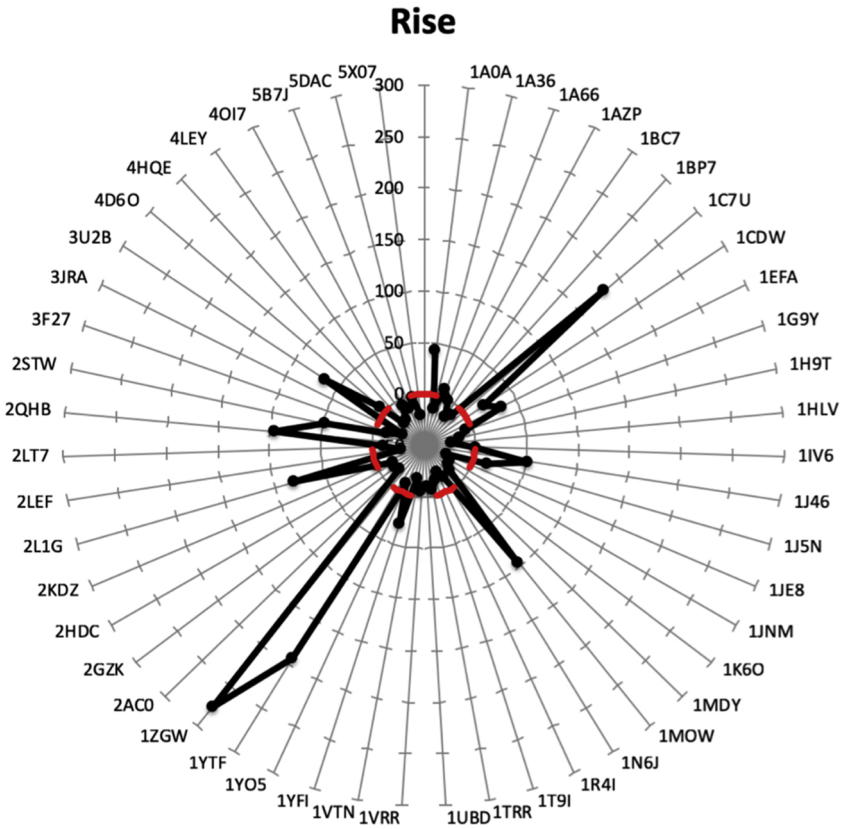

Rise
5X07
1A0A
1A36
1A66
1AZP
1BC7
1BP7
1C7U
1CDW
1EFA
1G9Y
1H9T
1HLV
1IV6
1J46
1J5N
1JE8
1JNM
1K6O
1MDY
1MOW
1N6J
1R4I
1T9I
1TRR
1UBD
1VRR
1VTN
1YFI
1YO5
1YTF
1ZGW
2AC0
2GZK
2HDC
2KDZ
2L1G
2LEF
2LT7
2QHB
2STW
3F27
3JRA
3U2B
4D6O
4HQE
4LEY
4OI7
5B7J
5DAC
300
250
200
150
100
50
0

Figure 2


(a)
HIS5,
LYS6,
ARG2,
ARG12-13
ARG13
GLU9
HIS94
ARG2, HIS5,
ARG12,GLU9,
ARG13, LYS6,
ASN73
LYS1
A
G
T
G
T
G
C
A
C
C
C
T
G
A
T
alpha
beta
chi
epsil
gamma
phase
zeta
0
2
4
6
Δ(bound-MDnaked)

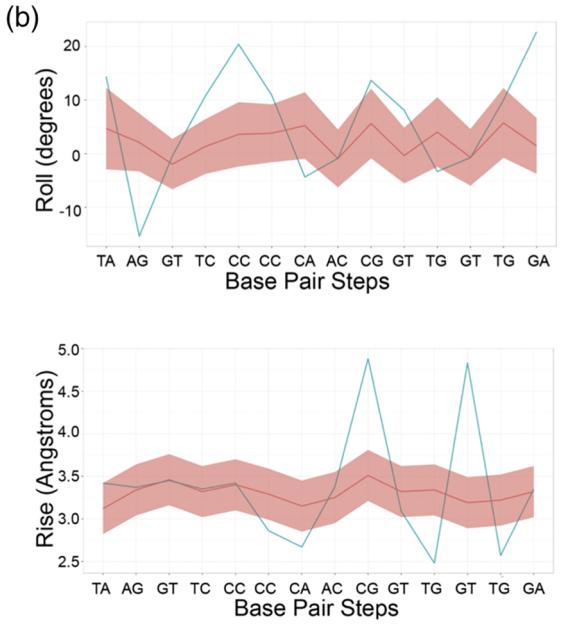

(b)
Roll (degrees)
20
10
0
-10
TA AG GT TC CC CC CA AC CG GT TG GT TG GA
Base Pair Steps
Rise (Angstroms)
5.0
4.5
4.0
3.5
3.0
2.5
TA AG GT TC CC CC CA AC CG GT TG GT TG GA
Base Pair Steps

Figure 3

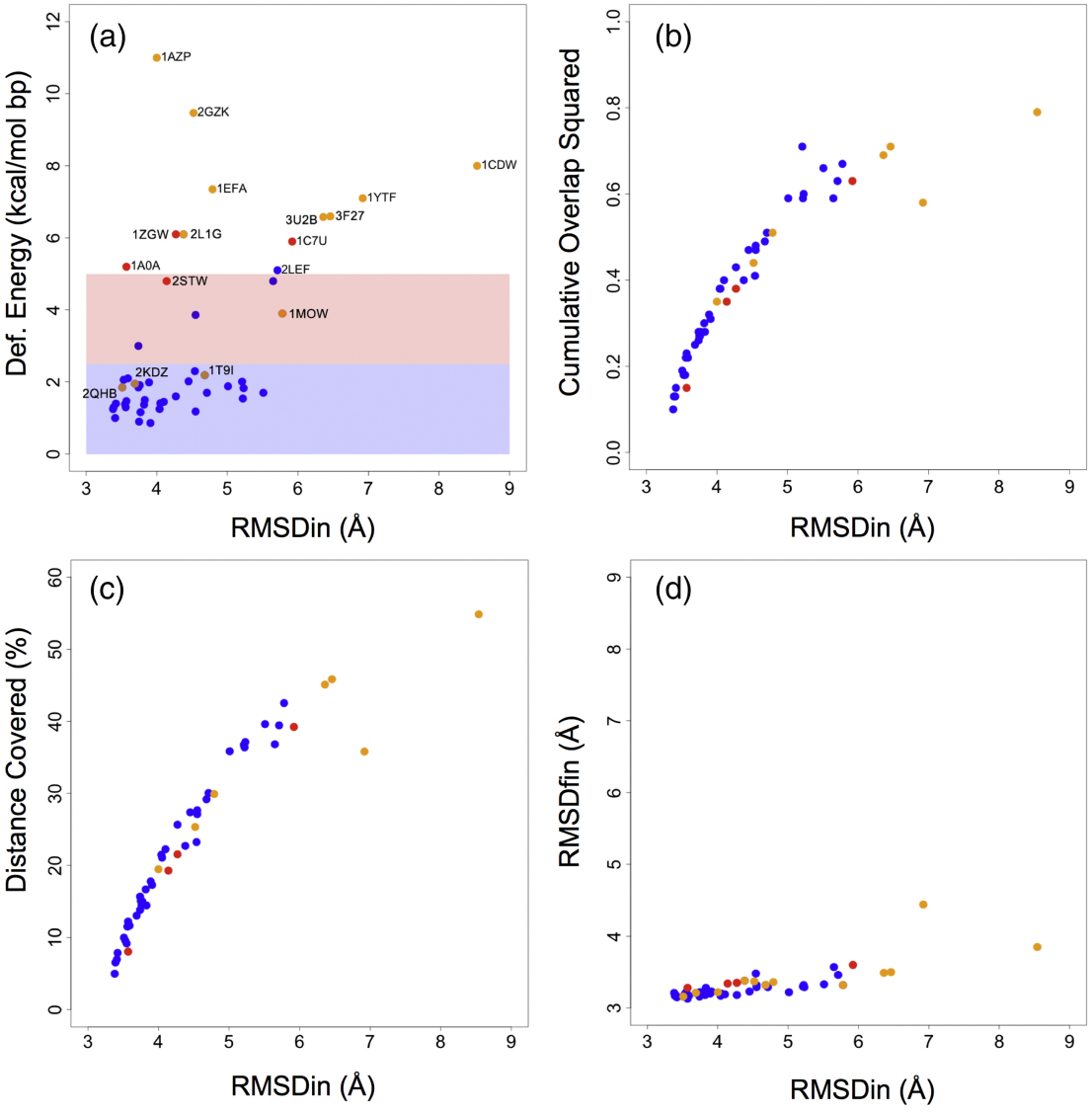

(a)
Def. Energy (kcal/mol bp)
RMSDin (Å)
1AZP
2GZK
1CDW
1EFA
1YTF
3U2B
3F27
1ZGW
2L1G
1C7U
1A0A
2LEF
2STW
1MOW
2KDZ
1T9I
2QHB
(b)
Cumulative Overlap Squared
RMSDin (Å)
(c)
Distance Covered (%)
RMSDin (Å)
(d)
RMSDfin (Å)
RMSDin (Å)

Figure 4

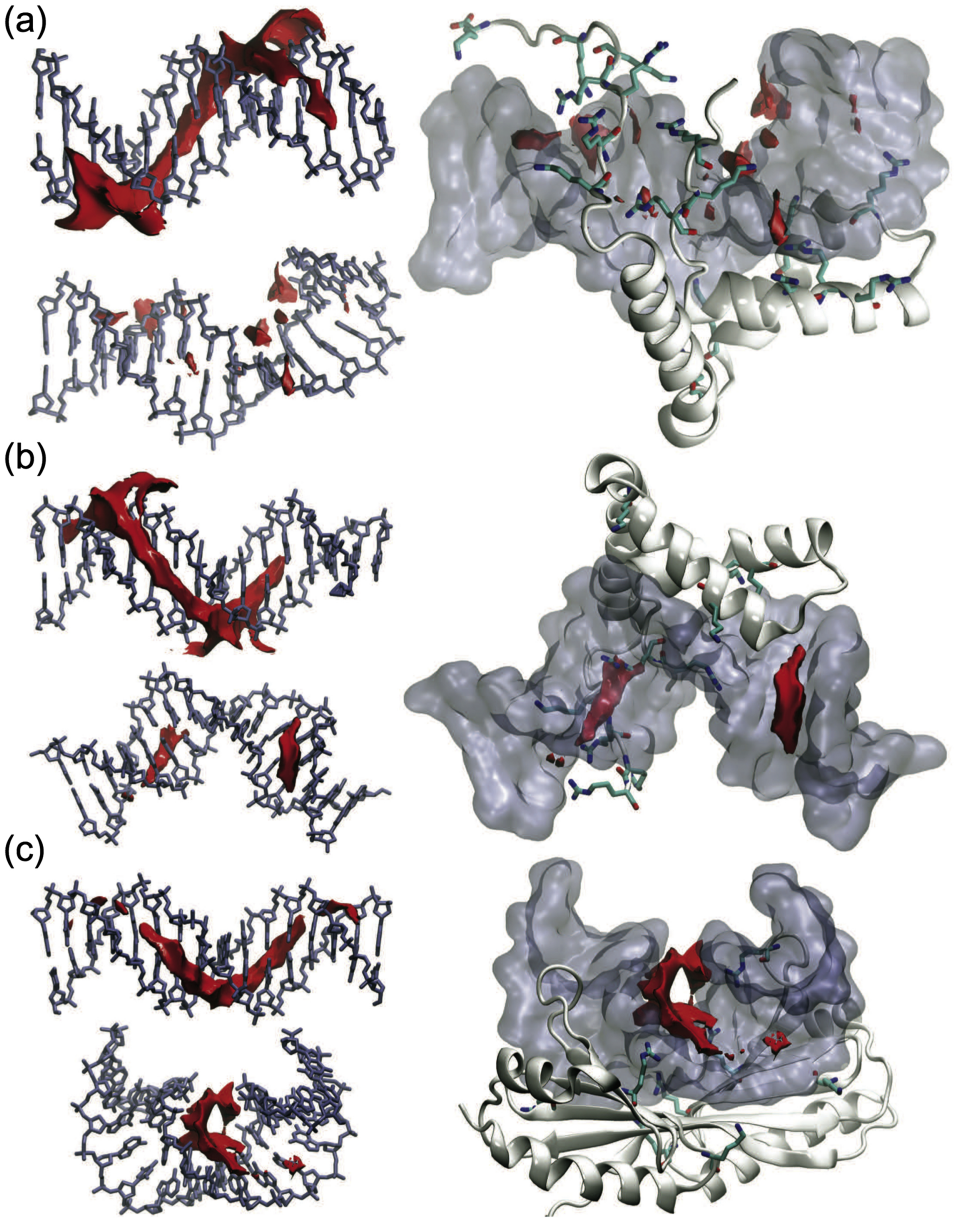


Figure 5

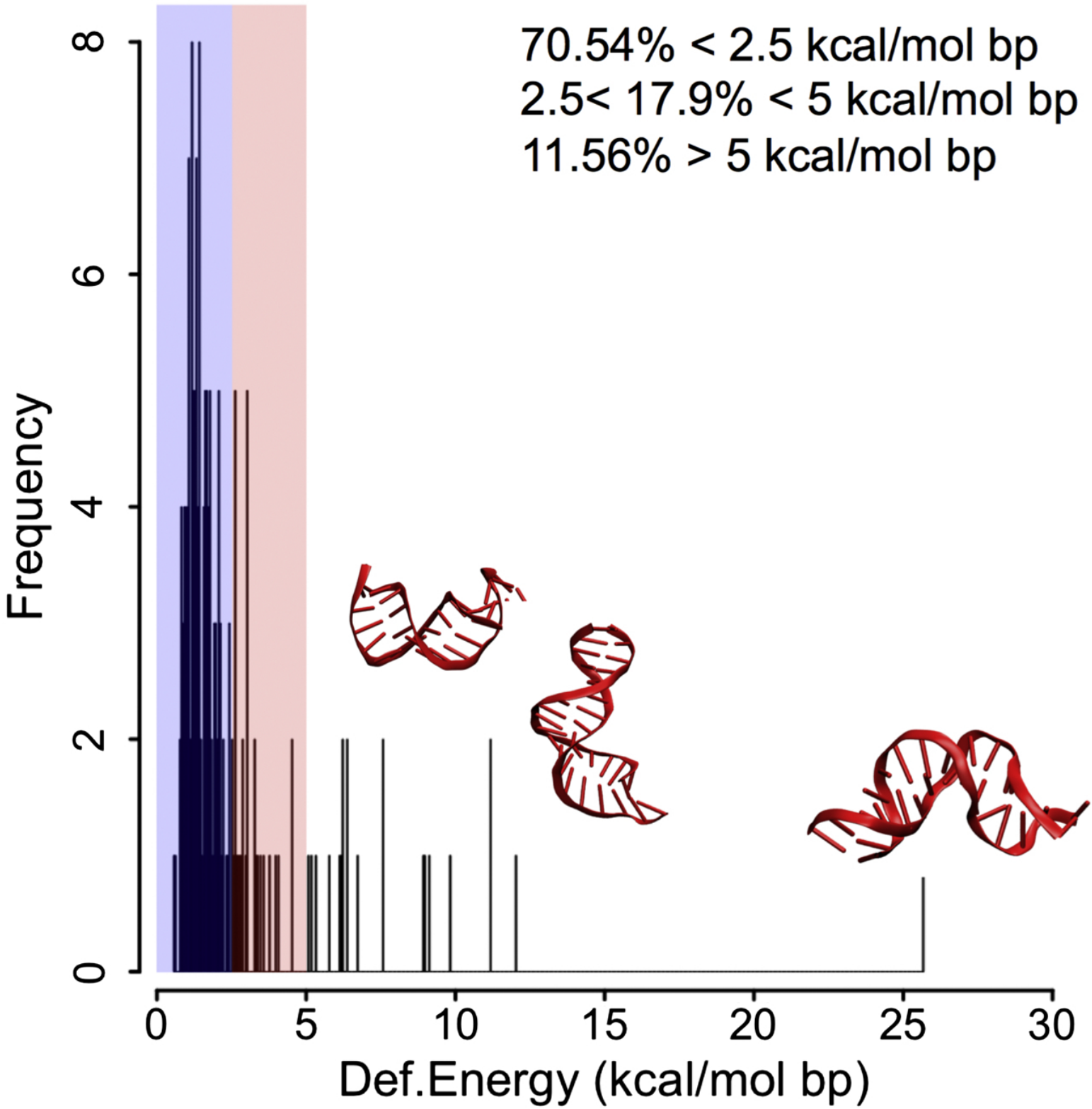

70.54% < 2.5 kcal/mol bp
2.5< 17.9% < 5 kcal/mol bp
11.56% > 5 kcal/mol bp
Frequency
0
2
4
6
8
0
5
10
15
20
25
30
Def.Energy (kcal/mol bp)

Figure 6

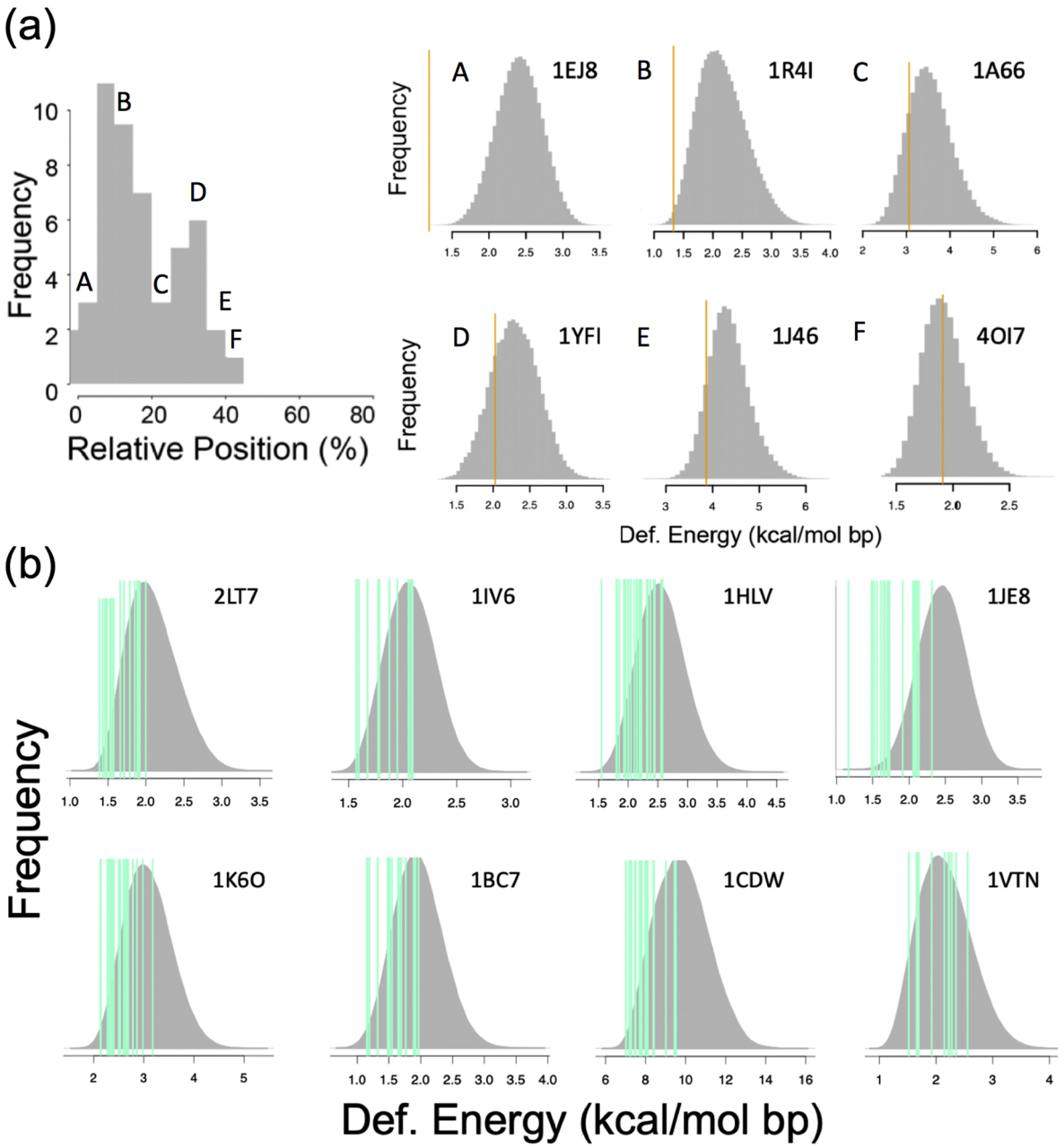

(a)
Frequency
Relative Position (%)
A
B
C
D
E
F
1EJ8
1R4I
1A66
1YFI
1J46
4OI7
Def. Energy (kcal/mol bp)
(b)
2LT7
1IV6
1HLV
1JE8
1K6O
1BC7
1CDW
1VTN
Frequency
Def. Energy (kcal/mol bp)

Figure 7